
PROYECTO: ANÁLISIS FILOGENÉTICO DEL SARS

GENÓMICA COMPUTACIONAL

ALUMNA:

KARLA ADRIANA ESQUIVEL GUZMÁN



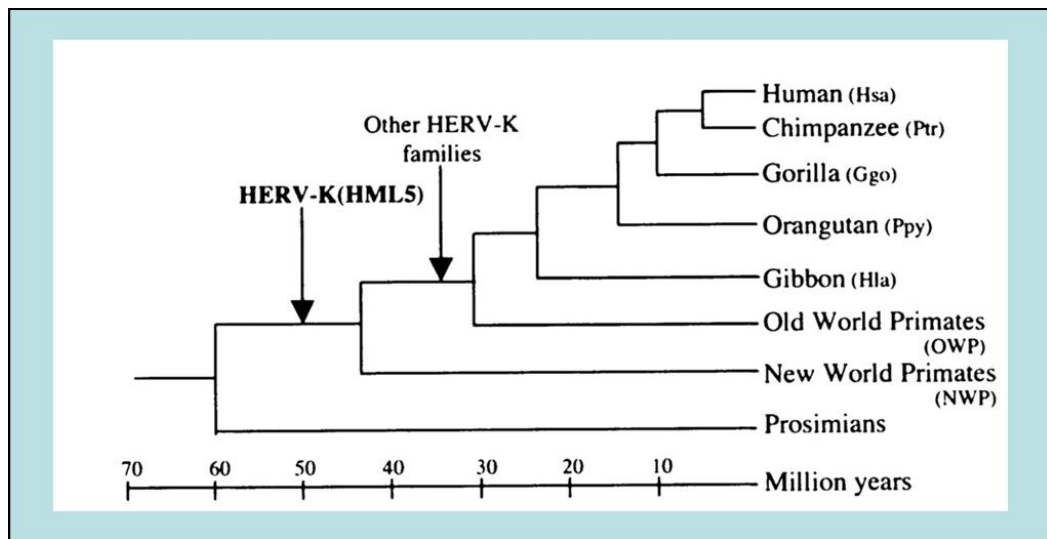
UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

13/DICIEMBRE/2018

Introducción

¿Qué es un árbol Filogenético? Los árboles filogenéticos representan una hipótesis acerca de las relaciones evolutivas entre grupos de organismos, un árbol filogenético puede construirse a partir de las características morfológicas, bioquímicas, conductuales o moleculares de las especies, básicamente cualquier ser vivo sobre la tierra puede rastrear su ascendencia a un ancestro en común.

A continuación un ejemplo de Árbol filogenético de los primates basado en comparaciones de la secuencia del retrovirus endógeno HERV-K(HML-5). Figura tomada de Lavie et al. 2004:



En este proyecto se va a modelar el análisis filogenético del SARS (Síndrome respiratorio agudo grave/Severe acute respiratory syndrome), es una neumonía atípica que apareció por primera vez en noviembre de 2002 en la provincia de Cantón, China. Se propagó a las vecinas Hong Kong y Vietnam a finales de febrero de 2003, y luego a otros países a través de viajes por medio aéreo o terrestre de personas infectadas. La enfermedad ha tenido una tasa promedio de mortalidad global cercana a un 13%. El 28 de febrero de 2003, el hospital francés de Vietnam en Hanoi, un hospital privado con

solo 60 camas, llamó a la Organización Mundial de la Salud (OMS) con un informe de pacientes que tenían síntomas inusuales similares a la influenza. Los funcionarios del hospital habían visto pasar un virus de influenza aviar a través de la región unos años antes y sospechaban de un virus similar. Principalmente me basé en el libro **Introduction to Computational Genomics** y el algoritmo principal que menciona es Neighbor Joining, sin embargo en este proyecto también se hace el cálculo con el algoritmo UPGMA para comparar, pero en realidad no hay gran diferencia.

- **Objetivo:** El objetivo del proyecto es mostrar el árbol filogenético del SARS, el cuál nos ayuda a conocer la evolución desde su primera aparición, la implementación de este proyecto se hizo en R.
- **Métodos:** Se utilizaron 2 algoritmos para calcular el árbol filogenético:
 - UPGMA (Unweighted Pair Group Method with Arithmetic Mean)
 - Neighbor joining

En ambos algoritmos utilicé los modelos “JC69” y “F81” para calcular distancias entre las secuencias de ADN. Las cadenas de ADN utilizadas las obtuve de GenBank <https://www.ncbi.nlm.nih.gov/genbank/>. Se buscan las secuencias de ADN con el accession number.

SARS-A POST-GENOMIC EPIDEMIC: PHYLOGENETIC ANALYSIS

Table 7.1 Name, location, and sampling date of SARS virus isolates used in our case study

Name of isolate	Acc. number	Date	Location
GZ01	AY278489	DEC-16-2002	Guangzhou (Guangdong)
ZS-A	AY394997	DEC-26-2002	Zhongshan (Guangdong)
ZS-C	AY395004	JAN-04-2003	Zhongshan (Guangdong)
GZ-B	AY394978	JAN-24-2003	Guangzhou (Guangdong)
HZS-2A	AY394983	JAN-31-2003	Guangzhou Hospital
GZ-50	AY304495	FEB-18-2002	Guangzhou (Guangdong)
CUHK-WI	AY278554	FEB-21-2003	Hong Kong
Urbani	AY278741	FEB-26-2003	Hanoi
Tor 2	AY274119	FEB-27-2003	Toronto
Sin2500	AY283794	MAR-01-2003	Singapore
TWI	AY291451	MAR-08-2003	Taiwan
CUHK-AG01	AY345986	MAR-19-2003	Hong Kong
CUHK-L	AY394999	MAY-15-2003	Hong Kong

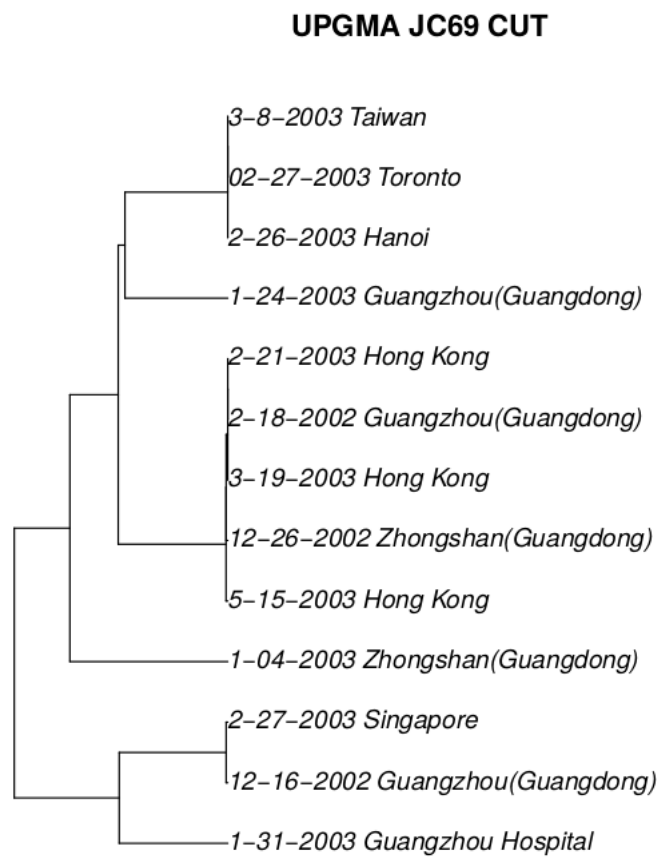
Nota: Las secuencias de ADN se descargaron en formato .fasta, tuvieron que ajustarse los tamaños de las secuencias pues R pide que sean del mismo tamaño para poder calcular las distancias y generar el árbol, por ello se encuentran 3 carpetas:

- ALL: se encuentran los archivos originales con las secuencias de ADN que se descargaron de GenBank.
- ALL-CUT: se encuentran las secuencias cortadas al tamaño de la secuencia de menor tamaño.
- ALL-FILL: se encuentran las secuencias rellenadas con “—” al tamaño de la secuencia de mayor tamaño.

Se ejecutó el algoritmo con cada uno de los ajustes ya mencionados.

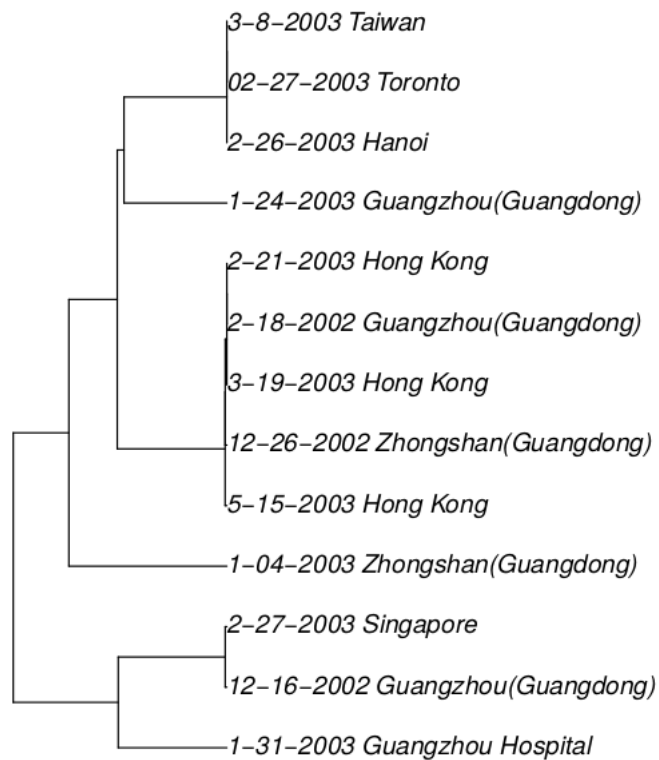
Resultados

Este árbol filogenético obtuvo con el algoritmo UPGMA utilizando el modelo para calcular distancias JC69 con la secuencia de ADN más corta:



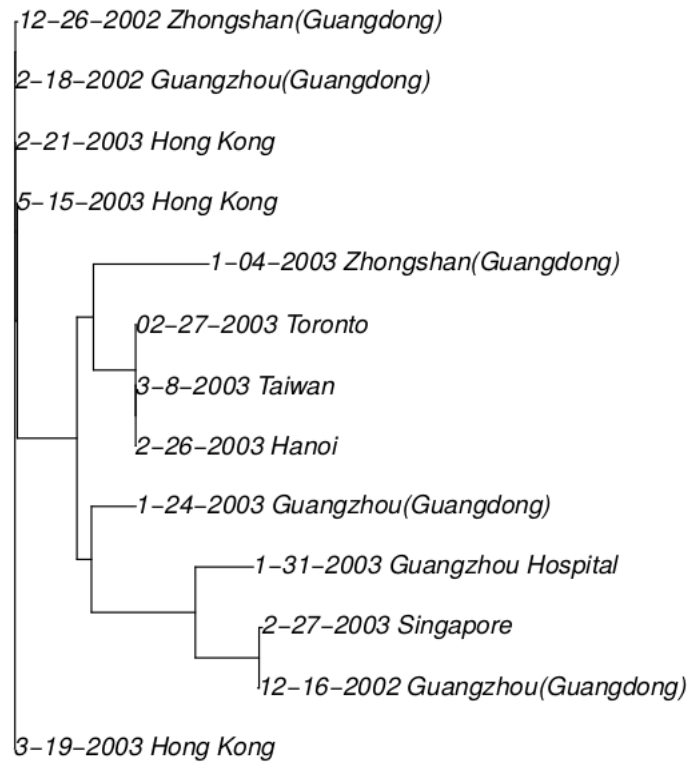
Este árbol se obtuvo con el algoritmo UPGMA utilizando el modelo para calcular las distancias F81 con la secuencia de ADN más corta:

UPGMA F81 CUT



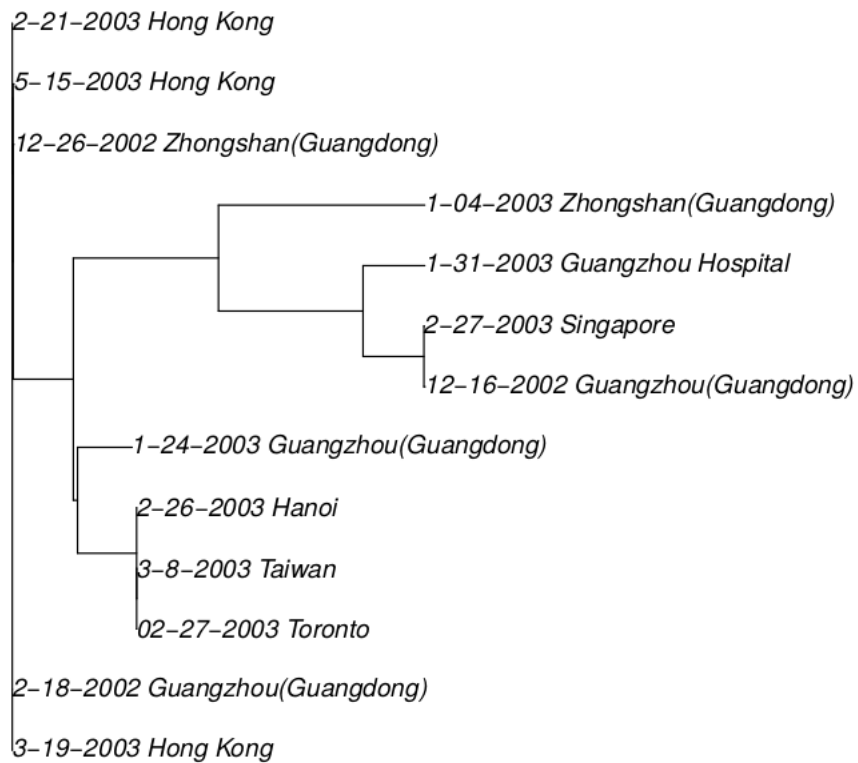
Este árbol se obtuvo con el algoritmo Neighbor Joining con el modelo para calcular distancias JC69 con la secuencia de ADN más corta:

Neighbor Joining JC69 CUT



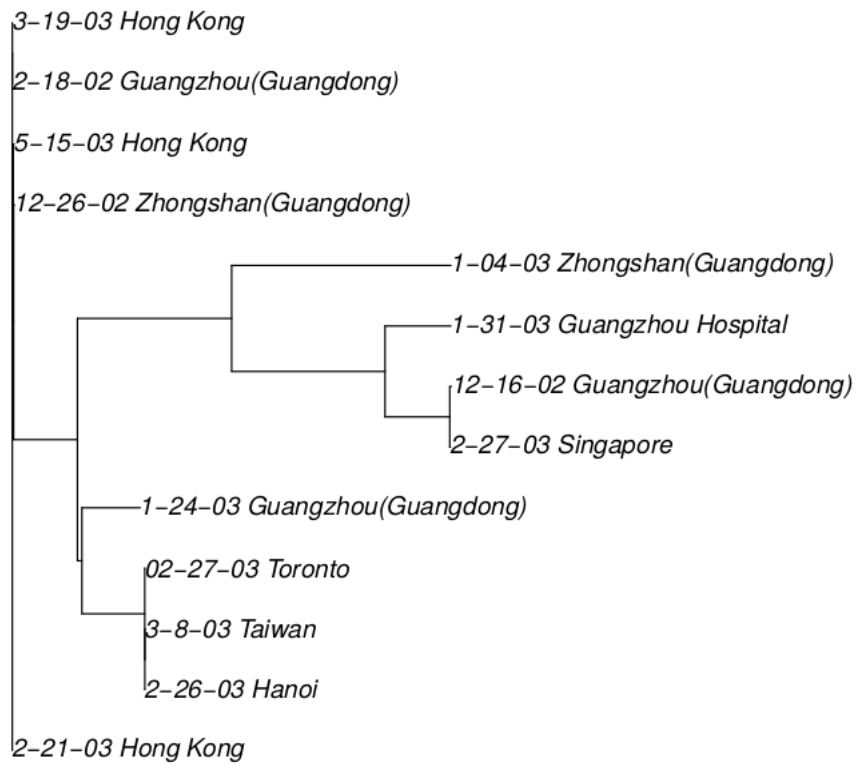
Este árbol se obtuvo con el algoritmo Neighbor Joining con el modelo para calcular distancias F81 con la secuencia de ADN más corta:

Neighbor Joining F81 CUT



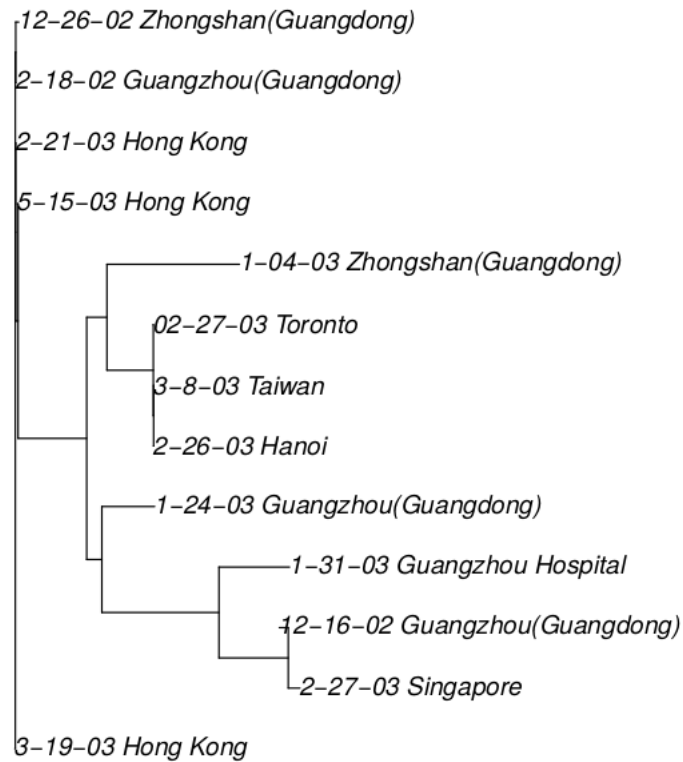
Este árbol se obtuvo con el algoritmo Neighbor Joining con el modelo para calcular distancias F81 con la secuencia de ADN más larga.

Neighbor Joining F81 FILL



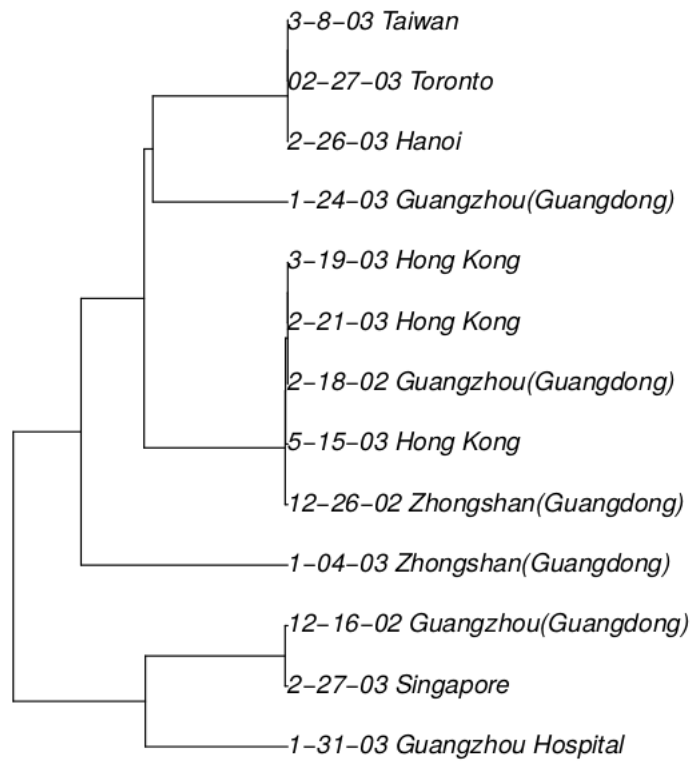
Este árbol se obtuvo con el algoritmo Neighbor Joining con el modelo para calcular distancias JC69 con la secuencia de ADN más larga.

Neighbor Joining JC69 FILL



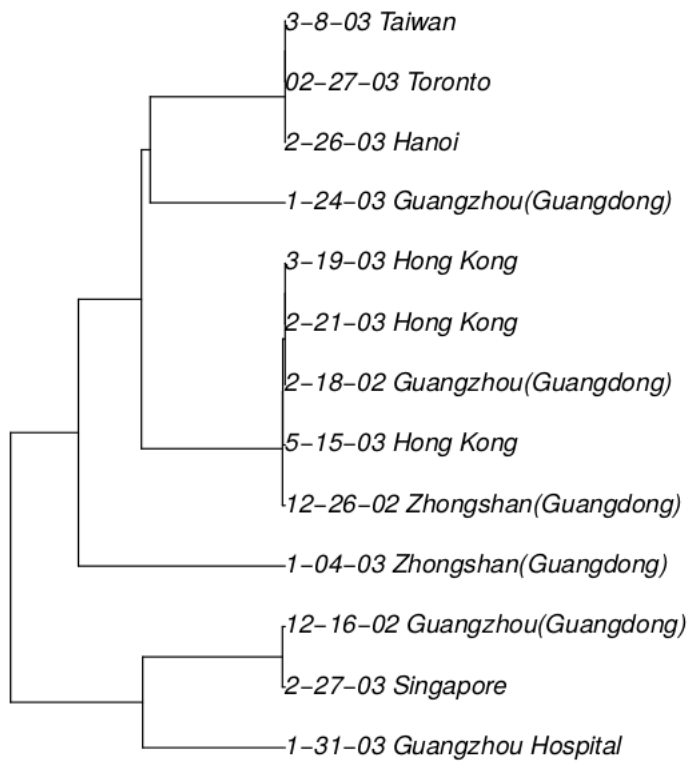
Este árbol se obtuvo con el algoritmo UPGMA con el modelo para calcular distancias F81 con la secuencia de ADN más larga.

UPGMA F81 FILL



Este árbol se obtuvo con el algoritmo UPGMA con el modelo para calcular distancias JC69 con la secuencia de ADN más larga.

UPGMA JC69 FILL



Discusión

Neighbor Joining funciona de la siguiente forma:

1. Basándose en la matriz de distancias actual calcula la matriz Q .
2. A continuación, busca el par de taxones para los que $Q(i, j)$ tiene su valor más bajo. Estos taxones se unen para formar un nuevo nodo que se une al resto del árbol.
3. Se calcula la distancia desde cada uno de los taxones del par a este nodo nuevo.
4. Se calcula la distancia desde cada uno de los taxones que no pertenecen a este nuevo par al nodo nuevo.
5. Se ejecuta el algoritmo otra vez, sustituyendo el par de taxones recién unidos por el nodo nuevo utilizando las distancias calculadas en el paso anterior.

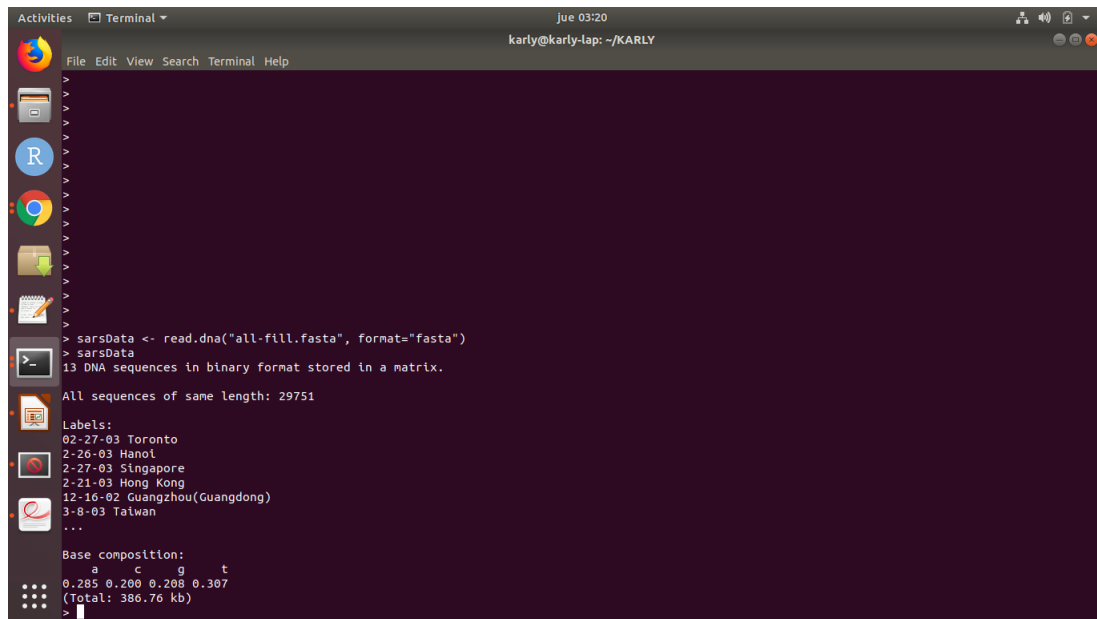
UPGMA (Unweighted Pair Group Method with Arithmetic Mean) funciona de la siguiente forma:

El algoritmo UPGMA construye un rooted tree (dendrograma) que refleja la estructura presente en una matriz de similitud de pares (o una matriz de disimilitud). En cada paso, los dos grupos más cercanos se combinan en un grupo de nivel superior. La distancia entre cualesquiera 2 clusters A y B cada uno con una cardinalidad de $|A|$ y $|B|$, se toma como el promedio de todas las distancias $d(x, y)$ entre pares de objetos $x \in A$ y $y \in B$, es decir, la distancia media entre los elementos de cada grupo. Requiere un supuesto de tasa constante, es decir, supone un árbol ultramétrico en el que las distancias desde la raíz hasta la punta de cada rama son iguales.

Sobre la implementación en R:

Esta línea es para leer las cadenas en formato .fasta:

```
sarsData <- read.dna("all-fill.fasta", format="fasta")
```

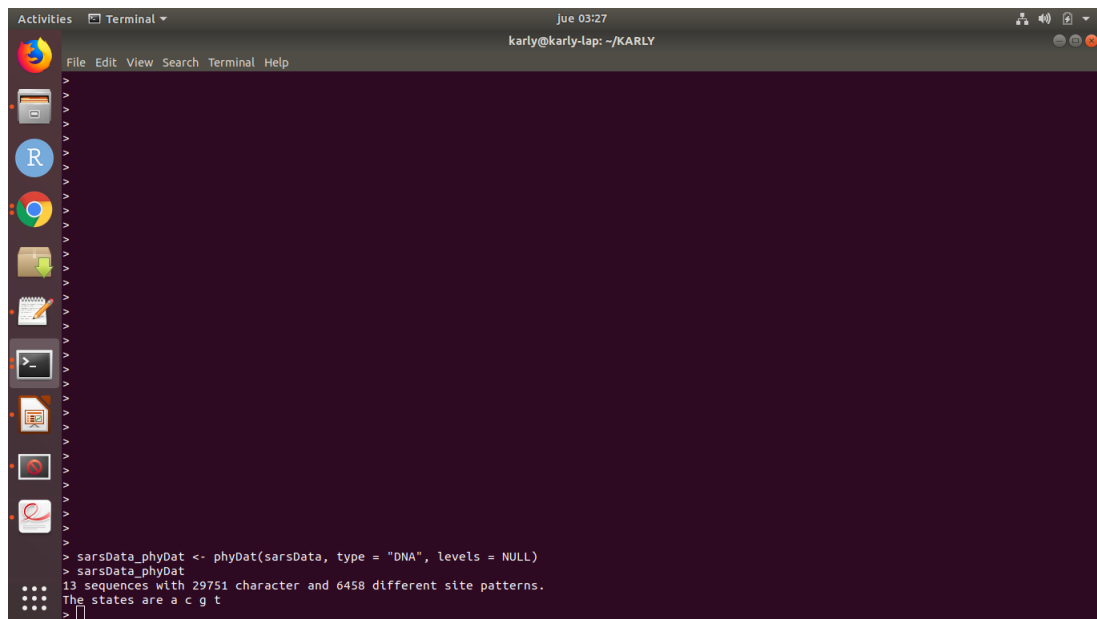


The screenshot shows a terminal window titled "Terminal" with a menu bar (File, Edit, View, Search, Terminal, Help) and a status bar (jue 03:20, karly@karly-lap: ~/KARLY). The terminal displays the following R code and output:

```
> sarsData <- read.dna("all-fill.fasta", format="fasta")
> sarsData
13 DNA sequences in binary format stored in a matrix.
All sequences of same length: 29751
Labels:
02-27-03 Toronto
2-26-03 Hanoi
2-27-03 Singapore
2-21-03 Hong Kong
12-16-02 Guangzhou(Guangdong)
3-8-03 Taiwan
...
Base composition:
      a      c      g      t
0.285 0.200 0.208 0.307
(Total: 386.76 kb)
>
```

Esta línea sirve para convertir al formato que R necesita para generar los árboles, nos dice cuantas secuencias y patrones se encontraron.

```
sarsData_phyDat <- phyDat(sarsData, type = "DNA", levels = NULL)
```



The screenshot shows a terminal window titled 'Terminal' with a menu bar (File, Edit, View, Search, Terminal, Help) and a status bar (Jue 03:27, karly@karly-lap: ~/KARLY). The terminal displays the following R commands and output:

```
> sarsData_phyDat <- phyDat(sarsData, type = "DNA", levels = NULL)
> sarsData_phyDat
13 sequences with 29751 character and 6458 different site patterns.
The states are a c g t
>
```

Esta linea hace la comparación de diferentes modelos de sustitución de nucleótidos o aminoácidos(calcula la distancia de ciertos patrones de aminoacidos).

```
mt <- modelTest(sarsData_phyDat)
print(mt)
```

```

> mt <- modelTest(sarsData_phyDat)
[1] "JC+I"
[1] "JC+G"
[1] "JC+G+I"
[1] "F81+I"
[1] "F81+G"
[1] "F81+G+I"
[1] "K80+I"
[1] "K80+G"
[1] "K80+G+I"
[1] "HKY+I"
[1] "HKY+G"
[1] "HKY+G+I"
[1] "SYM+I"
[1] "SYM+G"
[1] "SYM+G+I"
[1] "GTR+I"
[1] "GTR+G"
[1] "GTR+G+I"
> print(mt)
  Model df   logLik     AIC      AICw     AICc     AICcw     BIC
1    JC  23 -266271.6 532589.2 0.000000e+00 532589.2 0.000000e+00 532780.1
2   JC+I  24 -266267.7 532583.3 0.000000e+00 532583.3 0.000000e+00 532782.5
3   JC+G  24 -265000.2 530048.4 0.000000e+00 530048.5 0.000000e+00 530247.7
4 JC+G+I  25 -265000.7 530051.4 0.000000e+00 530051.4 0.000000e+00 530258.9
5    F81  26 -263259.7 526571.3 0.000000e+00 526571.4 0.000000e+00 526787.1
6   F81+I  27 -263257.1 526568.1 0.000000e+00 526568.2 0.000000e+00 526792.2
7   F81+G  27 -262062.3 524178.6 3.928904e-36 524178.6 3.970592e-36 524402.7
8 F81+G+I  28 -262062.7 524181.3 9.858815e-37 524181.4 9.944659e-37 524413.7
9    K80  24 -266270.4 532588.9 0.000000e+00 532588.9 0.000000e+00 532788.1
10 K80+I  25 -266266.7 532583.3 0.000000e+00 532583.3 0.000000e+00 532790.8
11 K80+G  25 -264998.4 530046.9 0.000000e+00 530046.9 0.000000e+00 530254.4
12 K80+G+I  26 -264998.9 530049.8 0.000000e+00 530049.8 0.000000e+00 530265.6
13   HKY  27 -263254.9 526563.9 0.000000e+00 526563.9 0.000000e+00 526788.0
14 HKY+I  28 -263252.4 526560.8 0.000000e+00 526560.9 0.000000e+00 526793.3
15 HKY+G  28 -262059.1 524174.3 3.386020e-35 524174.3 3.415504e-35 524406.7
16 HKY+G+I  29 -262059.5 524177.0 8.530625e-36 524177.1 8.588121e-36 524417.7
17    SYM  28 -266155.9 532367.7 0.000000e+00 532367.8 0.000000e+00 532600.1

```


En estas lineas se calculan las distancias con ambos Modelos JC69 y F81.

```
dna_distJC69 <- dist.ml(sarsData_phyDat, model="JC69")
```

```

> dna_distJC69 <- dist.ml(sarsData_phyDat, model="JC69")
> dna_distJC69
02-27-03 Toronto 2-26-03 Hanoi 2-27-03 Singapore
2-26-03 Hanoi 2.355131e-04
2-27-03 Singapore 5.465258e+00 5.516124e+00
2-21-03 Hong Kong 2.670287e+00 2.670893e+00 4.248408e+00
12-16-02 Guangzhou(Guangdong) 4.719644e+00 4.951972e+00 4.942048e-02
3-8-03 Taiwan 1.009183e-04 2.018638e-04 5.516124e+00
2-18-02 Guangzhou(Guangdong) 2.683010e+00 2.681807e+00 4.248408e+00
1-24-03 Guangzhou(Guangdong) 2.495317e+00 2.500031e+00 3.922556e+00
3-19-03 Hong Kong 2.675034e+00 2.681683e+00 4.248408e+00
12-26-02 Zhongshan(Guangdong) 2.671022e+00 2.669837e+00 1.000000e+01
1-31-03 Guangzhou Hospital 3.856298e+00 3.839220e+00 2.556195e+00
5-15-03 Hong Kong 2.685139e+00 2.683918e+00 4.961223e+00
1-04-03 Zhongshan(Guangdong) 3.339634e+00 3.333746e+00 4.875219e+00
2-26-03 Hanoi 2-21-03 Hong Kong 12-16-02 Guangzhou(Guangdong)
2-27-03 Singapore
2-21-03 Hong Kong
12-16-02 Guangzhou(Guangdong) 5.262523e+00
3-8-03 Taiwan 2.677541e+00 4.964524e+00
2-18-02 Guangzhou(Guangdong) 4.038770e-04 5.300591e+00
1-24-03 Guangzhou(Guangdong) 2.610930e+00 3.961358e+00
3-19-03 Hong Kong 4.036598e-04 5.262523e+00
12-26-02 Zhongshan(Guangdong) 4.889311e-02 4.682738e+00
1-31-03 Guangzhou Hospital 4.836541e+00 2.715395e+00
5-15-03 Hong Kong 4.265539e-02 5.493913e+00
1-04-03 Zhongshan(Guangdong) 4.217187e+00 4.025807e+00
3-8-03 Taiwan 2-18-02 Guangzhou(Guangdong)
2-26-03 Hanoi
2-27-03 Singapore
2-21-03 Hong Kong
12-16-02 Guangzhou(Guangdong)
3-8-03 Taiwan
2-18-02 Guangzhou(Guangdong) 2.684215e+00
1-24-03 Guangzhou(Guangdong) 2.497199e+00
3-19-03 Hong Kong 2.682335e+00 2.607648e+00
4.712110e-04

```

```
dna_distF81 <- dist.ml(sarsData_phyDat, model="F81")
```

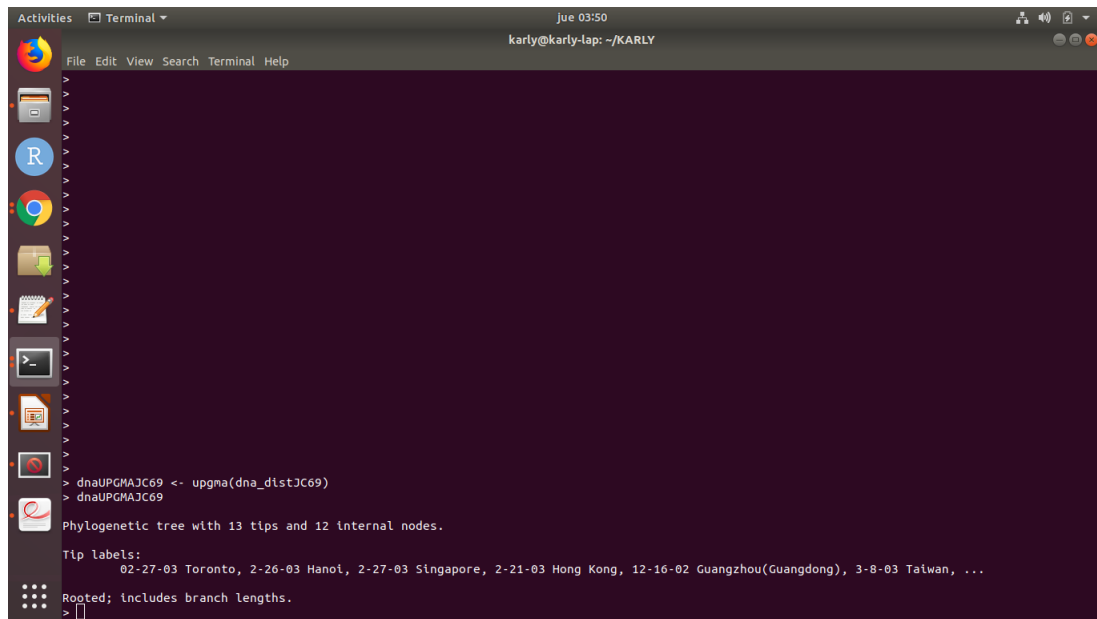
```

> dna_distF81 <- dist.ml(sarsData_phyDat, model="F81")
> dna_distF81
02-27-03 Toronto 2-26-03 Hanoi 2-27-03 Singapore
2-26-03 Hanoi 2.355158e-04
2-27-03 Singapore 1.000000e+01 1.000000e+01
2-21-03 Hong Kong 3.013645e+00 3.022976e+00 1.000000e+01
12-16-02 Guangzhou(Guangdong) 1.000000e+01 1.000000e+01 4.945375e-02
3-8-03 Taiwan 1.009196e-04 2.018658e-04 1.000000e+01
2-18-02 Guangzhou(Guangdong) 3.032759e+00 3.031119e+00 1.000000e+01
1-24-03 Guangzhou(Guangdong) 2.759171e+00 2.766060e+00 1.000000e+01
3-19-03 Hong Kong 3.022470e+00 3.031914e+00 1.000000e+01
12-26-02 Zhongshan(Guangdong) 3.023062e+00 3.021440e+00 1.000000e+01
1-31-03 Guangzhou Hospital 1.000000e+01 1.000000e+01 2.870593e+00
5-15-03 Hong Kong 3.043690e+00 3.042004e+00 1.000000e+01
1-04-03 Zhongshan(Guangdong) 1.000000e+01 1.000000e+01 1.000000e+01
2-26-03 Hanoi 2-21-03 Hong Kong 12-16-02 Guangzhou(Guangdong)
2-27-03 Singapore
2-21-03 Hong Kong
12-16-02 Guangzhou(Guangdong) 1.000000e+01
3-8-03 Taiwan 3.023899e+00 1.000000e+01
2-18-02 Guangzhou(Guangdong) 4.038770e-04 1.000000e+01
1-24-03 Guangzhou(Guangdong) 2.919254e+00 1.000000e+01
3-19-03 Hong Kong 4.036674e-04 1.000000e+01
12-26-02 Zhongshan(Guangdong) 4.892166e-02 1.000000e+01
1-31-03 Guangzhou Hospital 1.000000e+01 3.135409e+00
5-15-03 Hong Kong 4.268122e-02 1.000000e+01
1-04-03 Zhongshan(Guangdong) 1.000000e+01 1.000000e+01
3-8-03 Taiwan 2-18-02 Guangzhou(Guangdong)
2-26-03 Hanoi
2-27-03 Singapore
2-21-03 Hong Kong
12-16-02 Guangzhou(Guangdong)
3-8-03 Taiwan
2-18-02 Guangzhou(Guangdong) 3.034524e+00
1-24-03 Guangzhou(Guangdong) 2.762043e+00 2.914992e+00
3-19-03 Hong Kong 3.032838e+00 4.712137e-04

```

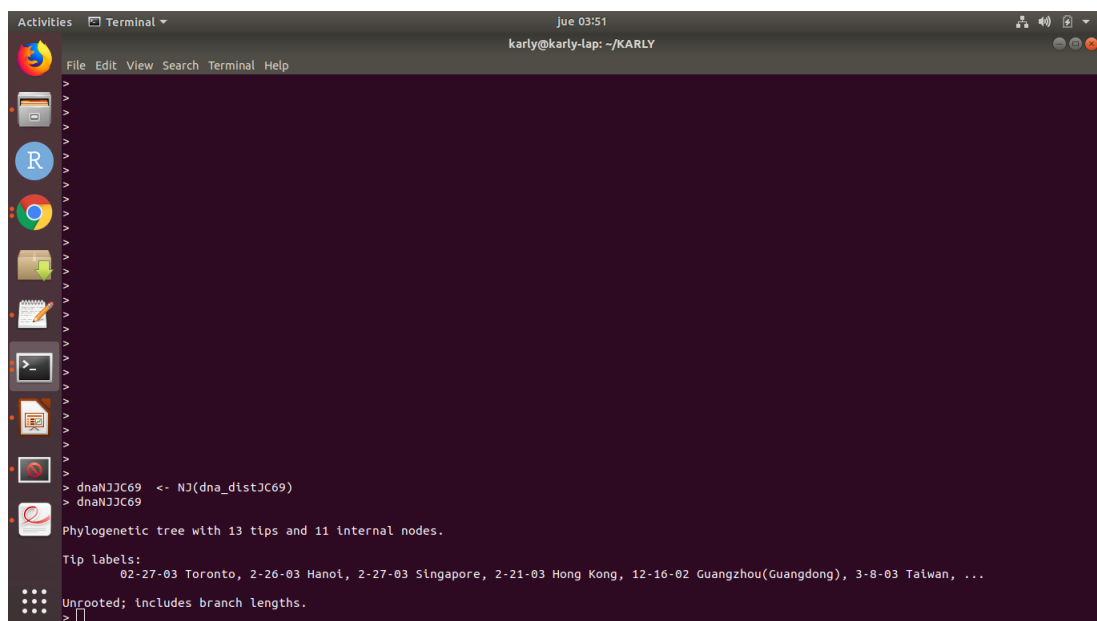
Estás líneas de código utilizan los algoritmos con cada uno de los modelos para calcular las distancias.

```
dnaUPGMAJC69 <- upgma(dna_distJC69)
```



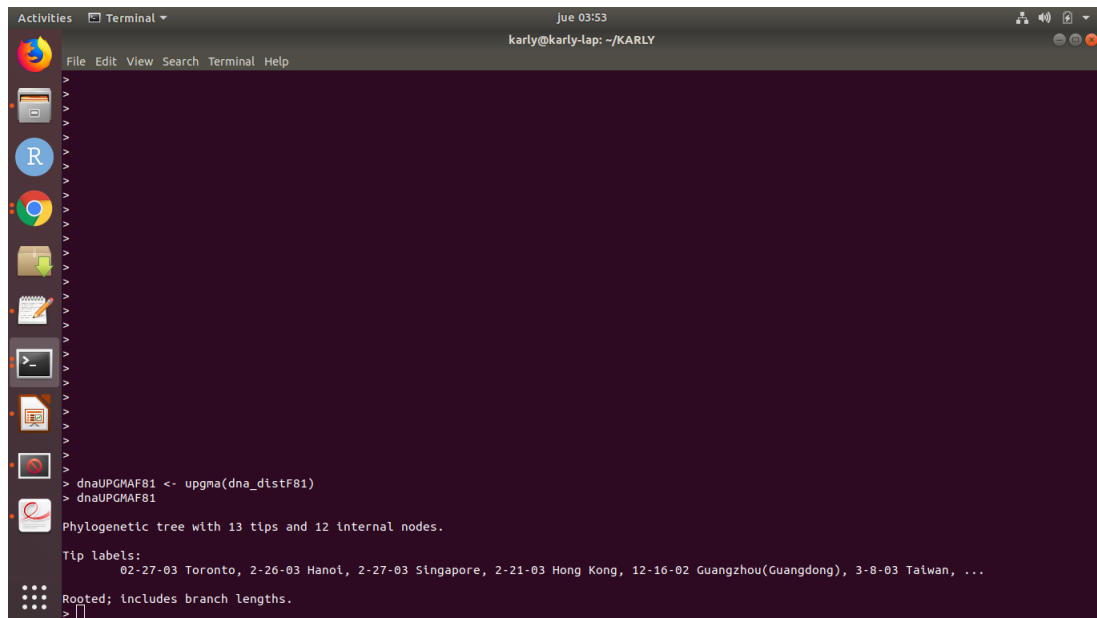
A terminal window titled 'Terminal' with a menu bar (File, Edit, View, Search, Terminal, Help) and a status bar (Jue 03:50, karly@karly-lap: ~/KARLY). The terminal shows the execution of R code: `> dnaUPGMAJC69 <- upgma(dna_distJC69)` and `> dnaUPGMAJC69`. The output is: `Phylogenetic tree with 13 tips and 12 internal nodes.` followed by `Tip labels:` and a list of locations and dates: `02-27-03 Toronto, 2-26-03 Hanoi, 2-27-03 Singapore, 2-21-03 Hong Kong, 12-16-02 Guangzhou(Guangdong), 3-8-03 Taiwan, ...`. It also shows `Rooted; includes branch lengths.` and a prompt `>`.

```
dnaNJJC69 <- NJ(dna_distJC69)
```



A terminal window titled 'Terminal' with a menu bar (File, Edit, View, Search, Terminal, Help) and a status bar (Jue 03:51, karly@karly-lap: ~/KARLY). The terminal shows the execution of R code: `> dnaNJJC69 <- NJ(dna_distJC69)` and `> dnaNJJC69`. The output is: `Phylogenetic tree with 13 tips and 11 internal nodes.` followed by `Tip labels:` and the same list of locations and dates as the previous screenshot. It also shows `Unrooted; includes branch lengths.` and a prompt `>`.

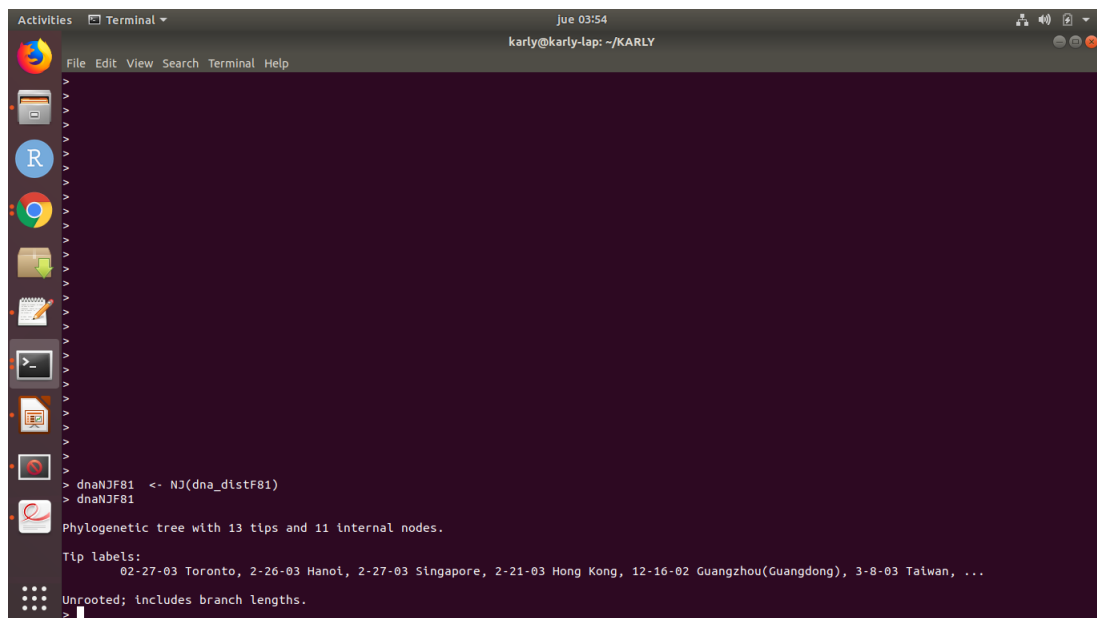
```
dnaUPGMAF81 <- upgma(dna_distF81)
```



A terminal window titled 'Terminal' with a menu bar (File, Edit, View, Search, Terminal, Help) and a status bar (jue 03:53, karly@karly-lap: ~/KARLY). The terminal shows the execution of R code: `dnaUPGMAF81 <- upgma(dna_distF81)` followed by `dnaUPGMAF81`. The output is: 'Phylogenetic tree with 13 tips and 12 internal nodes.' followed by 'Tip labels: 02-27-03 Toronto, 2-26-03 Hanoi, 2-27-03 Singapore, 2-21-03 Hong Kong, 12-16-02 Guangzhou(Guangdong), 3-8-03 Taiwan, ...' and 'Rooted; includes branch lengths.' The left sidebar shows various application icons including Firefox, R, Chrome, and a file manager.

```
> dnaUPGMAF81 <- upgma(dna_distF81)
> dnaUPGMAF81
Phylogenetic tree with 13 tips and 12 internal nodes.
Tip labels:
  02-27-03 Toronto, 2-26-03 Hanoi, 2-27-03 Singapore, 2-21-03 Hong Kong, 12-16-02 Guangzhou(Guangdong), 3-8-03 Taiwan, ...
Rooted; includes branch lengths.
>
```

```
dnaNJF81 <- NJ(dna_distF81)
```



A terminal window titled 'Terminal' with a menu bar (File, Edit, View, Search, Terminal, Help) and a status bar (jue 03:54, karly@karly-lap: ~/KARLY). The terminal shows the execution of R code: `dnaNJF81 <- NJ(dna_distF81)` followed by `dnaNJF81`. The output is: 'Phylogenetic tree with 13 tips and 11 internal nodes.' followed by 'Tip labels: 02-27-03 Toronto, 2-26-03 Hanoi, 2-27-03 Singapore, 2-21-03 Hong Kong, 12-16-02 Guangzhou(Guangdong), 3-8-03 Taiwan, ...' and 'Unrooted; includes branch lengths.' The left sidebar shows various application icons including Firefox, R, Chrome, and a file manager.

```
> dnaNJF81 <- NJ(dna_distF81)
> dnaNJF81
Phylogenetic tree with 13 tips and 11 internal nodes.
Tip labels:
  02-27-03 Toronto, 2-26-03 Hanoi, 2-27-03 Singapore, 2-21-03 Hong Kong, 12-16-02 Guangzhou(Guangdong), 3-8-03 Taiwan, ...
Unrooted; includes branch lengths.
>
```

Estás últimas 4 líneas sirven para graficar los árboles filogenéticos que se muestran en la sección de resultados.

```
plot(dnaUPGMAJC69, main = "UPGMA JC69 FILL")  
plot(dnaUPGMAJC69, main = "UPGMA F81 FILL")  
plot(dnaNJJC69, main = "Neighbor Joining JC69 FILL")  
plot(dnaNJF81, main = "Neighbor Joining F81 FILL")
```

Bibliography

- [1] Nello Cristianini y Matthew W. Hahn. Cambridge University Press, 2006.
- [2] https://en.wikipedia.org/wiki/Neighbor_joining
- [3] <https://en.wikipedia.org/wiki/UPGMA>